硕士学位论文

基于虚拟物理仿真思考的开放任务求解

PHYSICAL SIMULATION AND REASONING BASED TASK-AGNOSTIC LEARNING

柴士童

哈尔滨工业大学

2020 年 12 月

国内图书分类号：TM301.2　　学校代码：10213
国际图书分类号：62-5　　密级：公开

工程硕士学位论文

基于虚拟物理仿真思考的开放任务求解

硕士研究生：柴士童

导　　师：范晓鹏

申请学位：工程硕士

学　　科：软件工程

所在单位：计算学部

答辩日期：2020 年 12 月

授予学位单位：哈尔滨工业大学

Classified Index: TM301.2
U.D.C: 62-5

Dissertation for the Master's Degree in Technology

PHYSICAL SIMULATION AND REASONING BASED TASK-AGNOSTIC LEARNING

Candidate: CHAI Shitong
Supervisor: Professor FAN Xiaopeng
Academic Degree Applied for: Master of Technology
Specialty: Software Engineering
Affiliation: Faculty of Computing
Date of Defence: December, 2020
Degree-Conferring-Institution: Harbin Institute of Technology

摘　要

本文在Pyrobolearn机器人仿真环境中设计了一个要求机械臂的末端执行器到达指定物体附近的开放任务，并设计了新的强化学习算法用于训练智能体在未获得任务相关的奖励之前在未知环境中学习到可泛化到该开放任务的策略。

在 TD3 算法和 HER 算法的基础上，本文引入了基于局部敏感哈希和计数的奖励和基于正向动力学预测的奖励用于鼓励智能体探索未知环境。

本文提出的斯混合噪声层被用于提供自适应的策略噪声。提出的基于物理仿真引擎仿真时间的奖励被用于在未知任务的环境中鼓励智能体学习有效可泛化的策略，并在获得稀疏奖励的开放任务后快速适应到新的策略。

关键词：元学习；强化学习；机器人学；物理仿真；深度学习

Abstract

A task which requires the end effector of a manipulator to reach a specific body is designed with the help of the robot learning framework Pyrobolearn. New reinforcement learning algorithms are proposed to train an agent to learn in a task-agnostic environment without reward related to the task, where the learnt policy of the agent should be generalizable to the proposed task.

Based on TD3 and HER algorithms, a locality sensitive hashing based counting reward and a forward dynamics prediction model based reward are introduced to encourage the agent to explore in the unknown environment.

A mixed gaussian noise layer is proposed to provide a adaptive policy noise. A reward based on physics simulation time is proposed to encourage the agent to learn a generalizable policy in a task agnostic environment, and adapt to a new policy after given a task setting with sparse reward.

Keywords: Meta Learning, Reinforcement Learning, Robotics, Physics Simulation, Deep Learning

目　录

第 1 章 绪论

1.1 选题背景及研究意义

1.1.1 选题背景

一直以来，强化学习和机器人学都是人工智能研究中的热门领域。在 2008 年，Deepmind 团队基于深度强化学习研发的围棋人工智能系统 AlphaGo Zero 在零知识自我对弈的情况下在几天之内超越了旧的系统 AlphaGo，而 AlphaGo 曾击败了围棋领域中世界公认的专家柯洁等人[1]。这项研究使越来越多的人们开始关注人工智能领域，并使得强化学习成为研究热点。事实上，强化学习已经成为最热门的研究领域之一，并在自动控制、运筹学、机器人学、游戏智能体和无人驾驶等领域中获得了广泛的应用[2]。在这些领域中，机器人学是和前沿强化学习算法关系最密切的领域之一。传统的机器人如机械臂、四足机器人等，可以用强化学习训练得到的智能体进行控制，并在与环境交互过程中根据环境反馈使策略得到进一步的优化。

然而，由于机器人设计和制造的成本较高，通用的多关节机器人通常非常昂贵。而且机器人通常容易在强化学习中的各种随机探索中受到损坏，并导致控制系统和智能体策略在训练过程中出现错误。因此在实际的机器人上进行所有强化学习训练是不现实的[3,4]。为了避免这个问题，可以使用物理仿真引擎对机器人和环境进行建模，并在实际测试智能体之前先在仿真环境中对智能体进行训练。幸运的是，随着强化学习仿真需求的增加，越来越多的针对机器人的仿真环境开始出现，在这些仿真环境中，可以像在真实环境中一样控制机器人的关节、调节各种参数，或获得传感器数据等等，并可以做到在真实环境中难以做到的设定复杂的稀疏奖励、获取碰撞次数、和改变环境的物理参数等操作[5]。

虽然已经有大量的软件系统可以用于在一个环境建模完全精确的情形下解决一个良定义的任务[3]，如何让智能体在面对未知的新环境和未知的新任务后能够有效泛化之前学习到的策略仍然是一个未完全解决的问题。人类可以在陌生的环境中用很少次数的探索自然地掌握大量有效信息，还可以利用已有经验对大量物体进行分类、提高对物理实体运动的预测能力，或创新性地设计工具解决问题。由于人们对大脑的工作原理仍然知之甚少，这个过程通常很难被

数值化为一个单一的奖励函数或强化学习算法。

本课题致力于解决上述问题，即设计算法从而可以训练出能在任务奖励未知的环境中进行探索，获取环境信息，学习基本策略，并在任务确定后快速调节旧策略以适应新任务的智能体。为了实现这个目标，需要利用现有的开源物理仿真引擎、前沿的强化学习算法和具有强大函数拟合能力的深度神经网络来设计新算法。

1.1.2 研究意义

机器人控制对于工业制造有着重要的意义，在工厂流水线上，机械臂常常被设计为只能完成单个简单任务，或需要工人远程控制。虽然它们已经极大地提高了生产效率，减少了对工人们生命财产安全的威胁，但是机械臂由于价格昂贵，仅仅用于单一任务会造成极大的资源浪费。

本课题提出的算法有希望训练出可以在对未知任务奖励的环境进行充分探索之后，快速适应多种不同任务的机器人智能体，从而扩大现有强化学习算法的适用范围，解决更复杂的控制任务，增强机器人智能体泛化策略的能力。

此外，本课题还可以加深对现有强化学习算法在机器人控制中应用价值的理解，可以通过机器人仿真和控制帮助提前发现在应用算法到实际机器人控制时可能出现的问题，可以通过设计和调整深度神经网络进一步了解不同结构的神经网络对智能体性能的影响。

1.2 国内外研究进展

国内外已经有了很多关于让机器人智能体使用工具和泛化已学习的策略到新任务的研究。其中基于引导视觉的工具使用、无监督兴趣导向的探索、模仿学习和自驱动的主动学习与本课题有关。

1.2.1 基于引导视觉预见的工具使用

引导视觉预见[6] 可以使机械臂智能体从人类演示中学习并泛化学习到的能力以在不同的环境中使用工具。这种方法包含动作提案模型和预测模型。其中动作提案模型使用演示动作数据来训练一个自回归的长短时记忆网络模型来根据图像传感器拍摄到的图像数据生成动作。预测模型是基于卷积长短时记忆网络的[7]。预测模型被用于对物理实体的运动进行物理预测，并可以被用于筛选不能完成指定目标的动作序列。训练过程分为基于演示的模仿学习和利用握爪反射的随机自动训练。在测试过程中，指定的目标被定义为像素移动，预

测模型输出的像素位置和真实像素位置的距离被用于评估动作好坏。在指定任务后，动作序列从动作提案模型中采样，并使用交叉熵方法结合预测模型进行优化。实验表明，此种方法可以比单纯模仿学习在新环境中获得更好的泛化性能。

1.2.2 无监督兴趣导向探索

引导视觉预见需要有人类使用工具的演示数据才能正常工作，而对于更一般的环境，人类的演示数据可能是无法获得的，此时智能体应当可以在没有指定任务的情况下在环境中探索。为了解决这个问题，一个兴趣导向的探索方法被提出了[8]。这个方法结合了解纠缠目标空间的变分自编码机（VAE）和兴趣导向的 IMGEP（Intrinsically Motivated Goal Exploration Process）方法[9]。在目标空间被给定后，IMGEP 框架下的智能体会倾向于选择更有可能增加竞争力的目标。因为不像通常的强化学习算法一样在给定单一目标后做训练，而是无监督地选择目标进行训练，因此这是一个元策略算法。在探索过程中，智能体会学习到被 β-VAE 解纠缠后的观测，因此智能体可以分离地探索不同的物体。在实验中，此种方法获得了比单纯 IMGEP 方法更大的探索率。

1.2.3 模仿学习

在需要使用工具求解的开放任务中，往往有着稀疏奖励，随机探索很难刚好完成一个完整的动作序列，最终成功地使用工具完成任务，并获得奖励。模仿学习通过人类的专家策略，可以极大地加速这种随机探索过程。通过人类提供的演示数据，智能体应当能够学习到更好的探索策略，增大获得奖励的概率。相关研究表明，使用常规的强化学习算法和运动剪辑数据集，智能体可以学会组合学到的不同的技能，并用于求解多种任务[10]。不仅如此，智能体也可以从视频中学习到一些技能[11]。这意味着现有模仿学习方法可以利用网络中大量的视频数据进行学习。

1.2.4 自驱动的主动学习

在开放任务求解中，主动学习技术也可以被使用。为了在主动学习过程中获得预测物理环境的能力，可以使用带策略的循环 Q 网络来减少未知物理性质的熵[12]。在机器人学中，逆动力学模型对于稳健控制非常重要。一种主动学习方法使用竞争力来选择目标并在高维连续空间中学习各种技能，并用于机器人控制[13]。仿真结果表明这种方法能够帮助机器人智能体探索到随机策略难以探

索到的区域。

1.3 研究内容和方法

1.3.1 研究内容

在开放任务中，时刻 t 下会有一个稀疏奖励 $R_t \in \{0, -1\}$ 被提供给智能体。奖励为-1 表示任务失败，奖励为 0 表示任务成功。该奖励是关于环境状态的函数，即 $reward : \mathcal{S} \rightarrow \{0, -1\}$，其中 $\mathcal{S}$ 表示状态空间。智能体被允许在 $t < T_{start}$ 时间内对环境进行探索，但是无法获知函数 $reward$ 或由此函数计算得到的稀疏奖励 R_t，只能通过它观测到的状态向量 $s \in \mathcal{S}$ 来计算内部奖励（intrinsic reward），根据此奖励对环境进行探索，并学习能用于被 $reward$ 函数表示的任务有关的技能。在 $t \geq T_{start}$ 时，在每一时刻 t 对智能体提供由函数 $reward$ 计算得到的稀疏奖励 $R_t = reward(s)$。理想情况下，在 $t < T_{start}$ 时使用特定策略 π 进行探索的智能体应当能比使用随机选择动作的策略 π 进行探索的智能体更快地收敛到更高的奖励，即 $\mathbb{E}_{\pi}[R_t] > \mathbb{E}_{\pi'}[R_t]$。

1.3.2 研究方法

本文的实验环境是基于 Gym[14] 和 Pyrobolearn[15] 搭建的，其中 Gym 用于初步验证算法的正确性，大部分仿真实验在 Pyrobolearn 中完成。实验中对神经网络的优化使用了 Pytorch[16]，算法主体程序使用了 Python 语言。

本文的研究方法是基于内部奖励的强化学习。其中内部奖励被用于鼓励智能体在未知环境中探索，并掌握可用于解决开放任务的技能，它与任务特定的稀疏奖励 $reward$ 无关。

在本文的研究中，内部奖励与事后经验重放（Hindsight Experience Replay）被结合起来，用于获得更丰富的探索奖励。其中，多种内部奖励方法被使用，如基于局部敏感哈希的计数奖励、基于正向动力学预测的奖励和本文提出的物理仿真时间奖励。本文中的强化学习架构采用传统的演员 - 评论家（Actor-Critic）模式[17]，其中演员网络和评论家网络都使用了多层感知机。为了加速算法的收敛，减小训练过程中的不稳定性，训练算法借用了 TD3 算法[18] 的结构，引入了两个评论家，设计了目标网络以减缓演员网络和评论家网络参数更新的速度，并提出了混合高斯噪声层用于防止确定性策略收敛到局部极小。

第 2 章 相关算法基础

本文中提出的算法使用了大量最新强化学习算法中的思想，因此有必要介绍强化学习中常用的概念和现有强化学习算法的思想。

2.1 强化学习基础

在强化学习中，我们对智能体如何与环境交互感兴趣，这涉及到 4 个概念：模型、状态、动作和奖励[19]。由与智能体和环境相关的可观测量构成了智能体的状态向量 s，它表示了智能体当前在环境中所处的状态。每个状态从状态空间 $\mathcal{S}$ 中取值，$\mathcal{S}$ 包含了所有智能体和环境可以取到的状态。一个智能体可以在某一时刻做出动作 a 并因此转移到下一个状态。所有的动作从动作空间 $\mathcal{A}$ 中取值。如果完全已知一个环境的模型，且已知智能体的当前状态，就可以知道智能体做出任一动作的效果。当智能体做出某一个动作之后，根据当前的任务和环境可以得到一个奖励 r，它表示智能体从环境中得到的反馈，并从奖励空间 $\mathcal{R}$ 中取值。状态转换、动作和它导致的奖励构成了一个轨迹。如果用 S_t，A_t，R_t 分别表示在时刻 t 下智能体的状态、动作和奖励，那么轨迹就可以表示为一个序列 $S_t, A_t, R_{t+1}, S_{t+1}, \cdots, S_T$，其中 T 是关心的整个片断结束的时刻。策略 $\pi(a|s) = \mathbb{P}[A_t = a|S_t = s]$ 是智能体在当前状态 $S_t = s$ 时采取动作 $A_t = a$ 的条件概率。每个状态都有一个关联的值表示当前状态在特定任务下的价值，用状态价值函数 $V_\pi : \mathcal{S} \to \mathbb{R}$ 来表示从一个状态到这个值的映射。状态价值函数实际上表示了智能体使用当前策略 π 来选择动作，在可能获得奖励的多少。类似地，用动作价值函数 $Q_\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 表示智能体在给定当前状态和动作后，随后一直使用策略 π 来选择动作，在未来获得奖励的多少。在在实际算法中，为了防止对价值函数的拟合发散，需要给定一个折扣系数 $\gamma \in [0, 1]$ 来减少较远未来获得的奖励对当前价值函数的影响。强化学习的最终目标实际上为了最大化累积奖励：

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

有了累积奖励的定义后，状态价值函数 V_π 就可以写成在未来使用策略 π 的期望的累积奖励：

$$V_\pi(s) = \mathbb{E}_\pi[G_t|S_t = s]$$

类似地，动作价值函数可以写成：

$$Q_\pi(s,a) = \mathbb{E}_\pi[G_t|S_t = s, A_t = a]$$

2.2 TD3 算法

在通常的演员 - 评论家模式的强化学习算法中[17]，演员网络被用于拟合最优的确定性策略 $\mu^*: \mathcal{S} \to \mathcal{A}$，评论家网络被用于拟合最优的动作价值函数 $Q^*: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$。在给定一个迁移 (S_t, A_t, R_t, S_{t+1})，和现有的演员网络 μ 和评论家网络 Q 之后，根据如下损失函数对网络参数进行优化：

$$L_Q = \mathbb{E}[(R_t + \gamma Q(S_{t+1}, \mu(S_{t+1})) - Q(S_t, A_t))^2]$$

$$L_\mu = \mathbb{E}[-Q(S_t, \mu(S_t))]$$

但是由于在上述函数拟合过程中，评论家网络往往倾向于输出更高的动作值或状态价值。根据 TD3 算法[18]，可以使用两个评论家网络 $Q_1, Q_2: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$。这两个评论家网络同时对未来时刻 $t+1$ 时刻的动作和状态进行评估并分别输出价值 $Q_1(S_{t+1}, A_{t+1})$ 和 $Q_2(S_{t+1}, A_{t+1})$。取这两个值的最小值作为对未来时刻 $t+1$ 的价值预测，即可得到修正后的评论家网络的损失函数：

$$L_{Q_1} = \mathbb{E}[(R_t + \gamma \min\{Q_1(S_{t+1}, \mu(S_{t+1})), Q_2(S_{t+1}, \mu(S_{t+1}))\} - Q_1(S_t, A_t))^2]$$

$$L_{Q_2} = \mathbb{E}[(R_t + \gamma \min\{Q_1(S_{t+1}, \mu(S_{t+1})), Q_2(S_{t+1}, \mu(S_{t+1}))\} - Q_2(S_t, A_t))^2]$$

在对演员网络进行优化时，只使用评论家网络 Q_1：

$$L_\mu = \mathbb{E}[-Q_1(S_t, \mu(S_t))]$$

为了防止在训练评论家网络时出现发散或损失不稳定的情况，可以对每个网络引入一个靶网络。对每个靶网络权重，不使用损失函数对其进行优化，而是使用原网络和靶网络权重的指数滑动平均进行更新。对上述网络 μ, Q_1, Q_2，有对应的靶网络 μ', Q_1', Q_2'。在训练刚开始时，保持原网络和靶网络权重相同。在之后的训练过程中，使用上述损失函数对 μ, Q_1, Q_2 进行更新，而使用如下公式分别对靶网络的权重 $W_{\mu'}, W_{Q_1'}, W_{Q_2'}$ 进行更新：

$$W_{\mu'} = \tau W_\mu + (1-\tau) W_{\mu'}$$

$$W_{Q_1'} = \tau W_{Q_1} + (1-\tau) W_{Q_1'}$$

$$W_{Q_2'} = \tau W_{Q_2} + (1-\tau) W_{Q_2'}$$

其中 τ 叫做 polyak 系数，它控制着每次权重更新的多少。TD3 算法工作原理如图2-1所示，图中的替换使用了上述指数滑动平均方法，L_μ 表示演员网络的损失，L_{Q_1} 和 L_{Q_2} 分别表示根据上述公式计算出的评论家网络 1 和评论家网络 2 的损失。

2.3 事后经验重放算法

事后经验重放（HER）算法[20]是一个用于稀疏奖励的强化学习算法，它主要是为了解决奖励过少导致的训练速度过慢的问题。

事后经验重放要求状态向量 s 由观测向量 o，已完成目标 g^a，期望目标 g^d 构成，即 $s = (o, g^a, g^d)$。在智能体与环境交互的过程中，每当获得一个迁移 (s_t, a_t, r_t, s_{t+1}) 后，就把它放入重放经验缓冲（replay buffer）中。通常情况下，在训练时，从重放经验缓冲中直接取出一个个迁移并根据这些数据对网络进行训练。

而在事后经验重放算法中，为了增加成功奖励的个数，对每个取出的属于片段 e_i 的迁移 j，也即 (s_j, a_j, r_j, s_{j+1})，以一定概率 p_{future} 取出在它之后的一个迁移 k，也即 (s_k, a_k, r_k, s_{k+1})，其中有 $j \le k \le T$，T 是片段 e_i 的长度，并把迁移 j 的状态向量 $s_j = (o_j, g_j^a, g_j^d), s_{j+1} = (o_{j+1}, g_{j+1}^a, g_{j+1}^d)$ 中的期望目标 g_j^d, g_{j+1}^d 替换为未来在 k 时刻已完成的目标 g_k^a，最后根据新的状态对这些替换过的迁移重

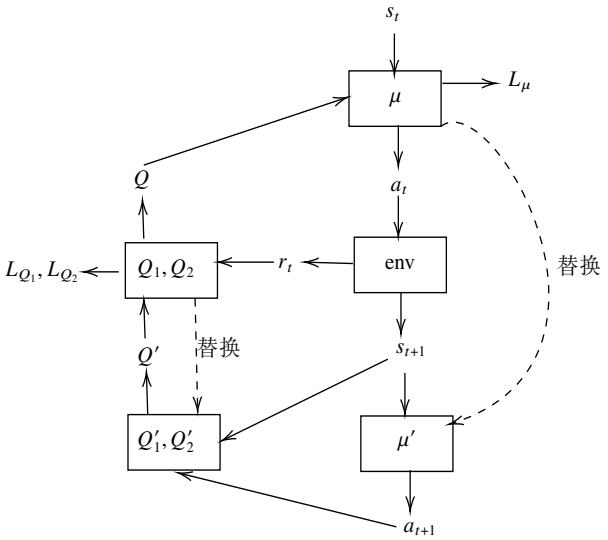

图 2-1 TD3 算法工作原理示意图

新计算奖励。

在上述算法中，概率 $p_{future} = 1 - \frac{1}{1+K_{replay}}$，其中 K_{replay} 是一个超参数，它决定了在经验重放时替换过的迁移所占比例的大小。

2.4 基于局部敏感哈希和计数的探索奖励

在开放任务中，由于奖励是稀疏的，智能体在自由探索过程中很难通过成功的奖励获知有关完成任务目标的信息，为了解决这一问题，需要引入任务无关的内部奖励来帮助智能体更快地探索有意义的状态，防止智能体对访问过的状态进行多次重复访问。

基于局部敏感哈希和计数的探索奖励[21] 使用一个局部敏感哈希函数对状态空间进行离散化，并对访问过的状态对应的哈希值进行计数，并通过此计数值计算出的奖励鼓励智能体访问之前未访问过的状态。

给定一个状态向量 $s \in \mathcal{S}$，可以使用 SimHash 函数计算它的哈希值：

$$\phi(s) = \mathrm{sgn}(As) \in \{-1, 1\}^k$$

其中 A 是随机生成的矩阵，满足：

$$A \in \mathbb{R}^{k \times \dim(\mathcal{S})}, A_{ij} \sim \mathcal{N}(0, 1)$$

A 在算法的初始化过程中随机生成，并在之后的训练过程中保持不变，而 k 是一个超参数，它的大小决定了哈希向量的长度，因此决定了离散化后的状态空间的粒度。

每当智能体到达一个状态 s 后，都可以根据上述哈希函数计算哈希值 $\phi(s)$，并使用一个字典对此哈希值进行计数，即：

$$freq[\phi(s)] += 1$$

因此字典 $freq$ 就记录了访问具有同样局部敏感哈希值的相似状态的频率，并可计算奖励：

$$reward_{lsh}(s) = \frac{1}{\sqrt{freq[\phi(s)]}}$$

2.5 基于正向动力学预测的探索奖励

每当智能体获得一个迁移 (s_t, a_t, r_t, s_{t+1})，可用于进行正向动力学预测的信息就更多了，这种信息蕴含在 $(s_t, a_t) \mapsto s_{t+1}$ 的映射中，它反映了环境的动力学

性质。

基于正向动力学预测的探索奖励[22]使用一个神经网络来拟合这种正向的动力学过程，即在理想情况下，预测模型 $f:\mathcal{S}\times\mathcal{A}\rightarrow\mathcal{S}$ 应当可以正确地预测未来的状态：$f(s_t,a_t)=s_{t+1}$。

对于经常访问到的简单状态和动作，正向动力学预测模型更有可能输出正确的预测值，而对于复杂未知的状态和动作，正向动力学预测模型则更有可能出错。这意味着可以把正向动力学预测模型的损失作为探索奖励来鼓励智能体探索更难被模型拟合的动力学过程。

给定一个迁移 (s_t,a_t,r_t,s_{t+1}) 和一个正向动力学预测模型 f，它的损失定义为：

$$L_{phy}=||f(s_t,a_t)-s_{t+1}||^2$$

根据以上分析，可以把这个损失值当作内部奖励提供给智能体，于是有奖励函数：

$$reward_{phy}(s_t,a_t,s_{t+1})=L_{phy}(s_t,a_t,s_{t+1})$$

第 3 章 实验环境

3.1 Gym 和 Pyrobolearn

Gym[14] 是一个由 OpenAI 团队开源的强化学习仿真环境，它已经成为了评价最先进的基线强化学习算法或机器人控制算法的标准环境。Mujoco[4] 是一个著名的商业物理仿真引擎，它在强化学习和机器人学中得到广泛应用，并成为 Gym 中机器人相关的环境的默认仿真引擎。在接下来的几章中，Gym 将被用来验证算法正确性，或被用于更方便地与其他现有算法进行对比。

Gym 提供一个 Python 语言的接口并且已经被包含在了 Python 包检索中。Gym 需要使用 OpenGL 的开源库 GLFW 来进行渲染。它可以直接使用 Python 包管理器 *pip* 进行安装。在 Gym 中，有很多内置的环境，包含大量著名的和强化学习有关基本任务可用于验证强化学习算法，并提供了统一的接口。例如，它包含一个经典的名为“推车杆”的控制任务，这个任务要求智能体通过水平地移动推车来平衡一个底部用自由活动的关节连接到推车上的杆。一个正常的强化学习算法应当可以在较长时间内避免杆失去平衡并倒下。

Mujoco（Multi-Joint dynamics with Cotact）是一个致力于仿真复杂关节运动、碰撞和多物体接触的物理引擎。它集成了粒子系统仿真、约束求解器、有限元积分器和凸优化器等等工具。它使用 ANSI C 编程并有一个 Python 的 API 包装。在 Gym 中，*FetchReach-v1* 环境需要使用 Mujoco 作为仿真器，因此在实验中 Mujoco、OpenGL 和 Gym 等工具都被安装在 Ubuntu 20.04 系统中，并设置了环境变量以保证正确的动态链接库 *libGL.so* 和 Mujoco 的二进制分发被加载。

一个 *FetchReach-v1* 环境的截图展示在图3-1中。 其中有一个机械臂被放在一张桌子前。在每个片段刚开始时，这个机械臂的状态都会被重置，而一个红点则会随机地出现在桌子上方的某处。机械臂智能体需要在片段的限制时间内使自己的末端执行器到达这个红点附近来完成任务并获得奖励。在片段的每一个时间步中，都会有观测值提供给此机械臂智能体。观测值包括指示关节状态、末端执行器位置坐标信息作为已完成的目标，和红点的位置坐标信息作为期望目标。指示关节状态的观测值 $o \in \mathbb{R}^{10}$ 是一个 10 维的向量，而末端执行器的坐标 $g^a \in \mathbb{R}^3$ 和红点的位置坐标 $g^d \in \mathbb{R}^3$ 都是 3 维的向量。因此提供给机械臂智能体的状态向量是一个 16 维的向量 $s = (o, g^a, g^d) \in \mathbb{R}^{16}$。如果末端执行器和红点

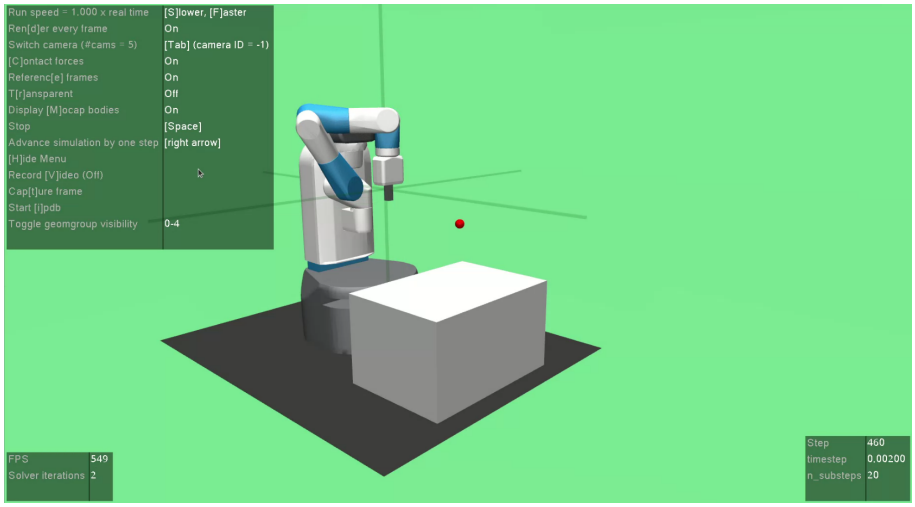

图 3-1 FetchReach-v1 环境的截图

之间的距离小于一个阈值，一个值为 0.0 的奖励会被提供给机械臂智能体，否则一个值为-1.0 的奖励会被提供给它。机械臂可以采取的动作是一个 4 维的取值在-1 到 1 之间的向量，即 $a \in [-1, 1]^4$。

3.2 Pyrobolearn 和 Pybullet

Pyrobolearn 是一个专门设计来训练智能机器人的框架[]。

3.3 环境参数

?

3.4 系统设计

系统由策略对象和奖励函数构成

第 4 章 混合高斯噪声层

4.1 网络结构

4.2 实验结果

第 5 章 仿真时间奖励

5.1 工作原理

5.2 实验结果

结　论

本文结合了 TD3 算法、HER 算法和基于 LSH 和计数的奖励、基于正向动力学预测的奖励，并提出了混合高斯噪声层和将物理仿真时间作为于未知任务的探索奖励。本文中的算法可以很好地用于训练可适应到给定开放任务的智能体，表现出与现有探索奖励相当的性能。

参考文献

[1] Silver D, Hubert T, Schrittwieser J, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play.[J]. Science, 2018, 362(6419) : 1140-1144.

[2] Dosovitskiy A, Ros G, Codevilla F, et al. CARLA: An Open Urban Driving Simulator[J]. Conference on Robot Learning, 2017 : 1-16.

[3] Toussaint M, Allen K R, Smith K A, et al. Differentiable Physics and Stable Modes for Tool-Use and Manipulation Planning.[C] // Robotics: Science and Systems XIV : Vol 14. 2018.

[4] Todorov E, Erez T, Tassa Y. MuJoCo: A physics engine for model-based control[C] // 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. 2012 : 5026-5033.

[5] Savva M, Malik J, Parikh D, et al. Habitat: A Platform for Embodied AI Research[C] // 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019 : 9339-9347.

[6] Xie A, Ebert F, Levine S, et al. Improvisation through Physical Understanding: Using Novel Objects As Tools with Visual Foresight[C] // Robotics: Science and Systems XV : Vol 15. 2019.

[7] Shi X, Chen Z, Wang H, et al. Convolutional LSTM Network: a machine learning approach for precipitation nowcasting[C] // NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 : Vol 28. 2015 : 802-810.

[8] Laversanne-Finot A, Péré A, Oudeyer P-Y. Curiosity Driven Exploration of Learned Disentangled Goal Spaces[C] // CoRL 2018 - Conference on Robot Learning. 2018 : 487-504.

[9] Forestier S, Mollard Y, Oudeyer P. Intrinsically Motivated Goal Exploration Processes with Automatic Curriculum Learning[J/OL]. CoRR, 2017, abs/1708.02190. http://arxiv.org/abs/1708.02190.

[10] Peng X B, Abbeel P, Levine S, et al. DeepMimic: example-guided deep reinforcement learning of physics-based character skills[J]. ACM Transactions on Graphics, 2018, 37(4) : 143.

[11] Peng X B, Kanazawa A, Malik J, et al. SFV: reinforcement learning of physical skills from videos[J]. ACM Transactions on Graphics, 2019, 37(6) : 178.

[12] Li S, Sun Y, Liu S, et al. Active physical inference via reinforcement learning.[J]. Cognitive Science, 2019 : 2126-2132.

[13] Baranes A, Oudeyer P-Y. Active learning of inverse models with intrinsically motivated goal exploration in robots[J]. Robotics and Autonomous Systems, 2013, 61(1) : 49-73.

[14] Brockman G, Cheung V, Pettersson L, et al. OpenAI Gym[J], 2016.

[15] Delhaisse B, Rozo L, Caldwell D G. PyRoboLearn: A Python Framework for Robot Learning Practitioners[C/OL] // Conference on Robot Learning (CoRL). 2019. https://github.com/robotlearn/pyrobolearn.

[16] Paszke A, Gross S, Massa F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library[C] // Advances in Neural Information Processing Systems : Vol 32. 2019 : 8026-8037.

[17] Konda V R, Tsitsiklis J N. Actor-critic algorithms[M]. 2002.

[18] Fujimoto S, van Hoof H, Meger D. Addressing Function Approximation Error in Actor-Critic Methods[J/OL]. CoRR, 2018, abs/1802.09477. http://arxiv.org/abs/1802.09477.

[19] Sutton R, Barto A. Reinforcement Learning: An Introduction[M]. 1988.

[20] Andrychowicz M, Wolski F, Ray A, et al. Hindsight Experience Replay[J/OL]. CoRR, 2017, abs/1707.01495. http://arxiv.org/abs/1707.01495.

[21] Tang H, Houthooft R, Foote D, et al. #Exploration: A Study of Count-Based Exploration for Deep Reinforcement Learning[J/OL]. CoRR, 2016, abs/1611.04717. http://arxiv.org/abs/1611.04717.

[22] Stadie B C, Levine S, Abbeel P. Incentivizing Exploration In Reinforcement Learning With Deep Predictive Models[J/OL]. CoRR, 2015, abs/1507.00814. http://arxiv.org/abs/1507.00814.

附录 A 算法代码

A.1 训练代码

A.2 测试代码

索引

哈尔滨工业大学学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《基于虚拟物理仿真思考的开放任务求解》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：　　日期：　　年　月　日

学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

致　谢

衷心感谢导师 范晓鹏 教授对本人的精心指导。他的言传身教将使我终生受益。

感谢哈工大 LaTeX 论文模板 hiThesis 。